Tailoring Conversational UX through the Lens of Dialogue Complexity

Q. Vera Liao
Biplav Srivastava
Pavan Kapanipathi
IBM T.J. Watson Research
1101 Kitchawan Rd,
Yorktown Heights, NY 10598 USA
vera.liao@ibm.com, {biplavs, kapanipa}@us.ibm.com

Abstract
Interactions with conversational systems take diverse forms across applications and individual users. In order to develop intelligent systems that can accommodate these diverse conversational user experience (UX) needs and preferences, meanwhile balancing implementation costs, we propose a computation-driven approach to profile conversational interactions by measuring dialogue complexity of user inputs in multiple dimensions. To inform system adaption designs, we propose to conduct comparative conversation analysis and user experiments, in order to develop classified UX guidelines for conversational interactions with different complexity profiles.

Author Keywords
Conversational UX; dialogue complexity; adaptive system; personalization.

ACM Classification Keywords
H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

Introduction
In recent years, we are seeing growing excitement around conversational systems or "chatbots", as many of them (e.g., Siri, Alexa) have become our daily encounters. This is not merely the rekindling of our longtime fascination with computer systems that talk like human. We are welcoming a shift from graphical user interface (GUI) to conversational interface, through which we interact with computer programs for various tasks such as search and navigation. A corresponding shift in the tasks of UX researchers is inevitable. For example, new analytical methods need to be introduced to derive design principles for lexical choices, sentence composition, discourse structure, and

high	low
U: *What is the payment schedule of payroll service?*	U: *When am I getting paid?*

Lexical complexity

high	low
U: *I'm trying to sign onto the wifi at Littleton. I can only sign onto the visitor wifi. It won't accept my password and username.*	U: *How can I sign onto wireless internet at a different work location?*

Information complexity

high	low
U: *How many vacation days do I have?* A: *Vacation time depends on your service time.* U: *I just joined this year.*	U: *how many days of vacations do I have if I joined in September this year?*

Structure complexity

high	low
U: *What can you help?* *A: I can help with information about IBM* *U: How do I sign up for healthcare benefits?* *A: To sign up for healthcare benefits, you can visit this website [link]* *U: Good job. Thank you!*	*U: healthcare benefits.* *A: To sign up for healthcare benefits, You can visit this website [link]*

Interaction complexity

Table 1: Examples of different conversational styles

so on, for which many suggest to learn from the disciplines of linguistics and conversation analysis (CA).

To embark on establishing conversational UX as a research discipline, we must take an inclusive view to consider its broad scope. In addition to the most hyped virtual assistants that handle a diverse set of natural conversations, the extended family of conversational systems also comprise some "mundane" members that talk like "English, enter 1", "reserved" ones that only answer factual questions on certain topics, "light-hearted" ones that aim to keep users entertained, and so on. These systems not only vary drastically in their knowledge, vocabulary, and affordance of interactions, but also their development and running costs.

Meanwhile, users choose to engage in different styles of conversational interactions, even with the same system. As an example, in our experience with deploying a Human Resource (HR) bot that answers questions from company employees [3], we observed a diversity of interaction styles that differ in complexity of various aspects (Table 1), including :1) *lexical complexity*, that some users tend to use more specialized or formal vocabulary; 2) *information complexity*, that some users have a tendency to give elaborate information when asking questions; 3) *structure complexity*, that some are more likely to go back-and-forth with the system in short turns; 4) *interaction complexity*, that some users enjoy diverse types of interactions beyond querying about functional information, such as social chit-chat.

To accommodate these individual differences, we envision conversational systems that can infer users' preferred interaction styles based on their input signals and tailor its system architecture and conversational designs accordingly. For instance, for users who tend to use simple keywords or monotonous form of input, it may be preferable and also cost-effective to implement it as a query-based or question-and-answer (QA) system. For those talking in a complex manner, to satisfy user needs, not only does the system need to be equipped with additional modules such as context and information state trackers, but it also needs to be designed with richer conversational structures to handle the dialogue flows.

Studying these differences in conversational styles and tailoring conversational UX is especially important with emerging technologies for "conversational platforms". The powerful future role of conversational interface may be for operating system that hosts and interconnects various applications, with which users interact through a bot persona. Under this assumption, we are likely to see even more diverse and frequently changing styles of conversational interactions. For example, when conversing about different domains or with different applications, there may also be complexity differences in *domain entities*, *semantics*, *ontologies*, and also, *interaction modalities*.

Dialogue Complexity

In this position paper, we offer a novel perspective on *profiling conversational interactions by* ***dialogue complexity*** *and developing classified UX guidelines based on the complexity profiles*. Aiming to enable intelligent systems that can automatically tailor conversational UX based on user input, there are two key considerations that distinguish our approach from previous works. The first one is *computation-driven*. While there can be many ways to classify conversational systems (e.g., by application domain,

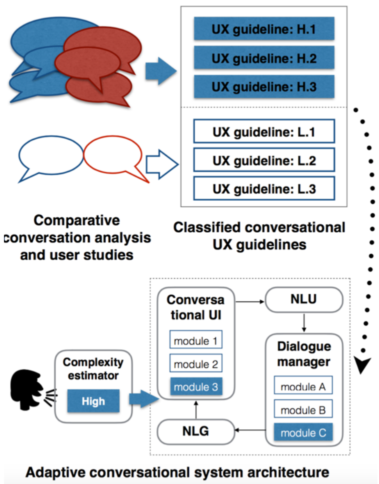

Figure 1: Overview for conducting comparative studies of conversational interactions with different complexity profiles to inform development of adaptive conversational systems.

system goals), we steer away from context-dependent taxonomy but attempt to develop a data-driven measure that is able to profile conversational interactions by the user input itself.

The second consideration is *multi-dimensionality*. In linguistics, dialogue complexity is traditionally addressed from a human readability point of view by identifying linguistic markers for "elaborated" styles (e.g. relative clause) [1]. To address "system readability" ---capability to handle complex conversational interactions, and balance implementation costs, we need to define complexity in multiple dimensions that have correspondence in specific modules of dialogue systems. For example, it would be more expensive to develop and maintain a system that handles more complex (e.g. specialized) *vocabulary*, or one that needs to recognize and track complex *domain entities and relations*.

In addition to answering *what to tailor* by defining the complexity measures, we need to understand *how to tailor* conversational UX to enable the system adaption. We propose to *conduct comparative studies of conversational interactions with different complexity profiles to derive classified UX guidelines*. We note that the methods of comparative studies should be orthogonal to the definition of complexity, and our goal is to compare across multiple complexity dimensions to develop a comprehensive set of UX guidelines. In the following section, we focus on proposing a general plan for the comparative studies to inform tailoring of conversational UX. At the end of the paper we briefly illustrate our idea with a multi-dimensional dialogue complexity scheme and desired results.

Study Overview

In Figure 1, we illustrate the motivation for conducting comparative studies of conversational interactions with different complexity profiles. The end goal, as shown at the bottom, is to inform development of adaptive system architecture. The general idea is that different modules of the system will be mapped to be used, or prioritized, for different levels of dialogue complexity, based on the output given by the complexity estimator. By "modules", we broadly refer to both ones in the back-end system such as various recognizers, trackers and planners, and also "conversational modules" equipped in the interface such as repair, explanation, and affective dialogues. The mapping relations will be informed by the classified UX guidelines empirically derived from the comparative studies.

As an example, we propose to compare interactions with simple and complex discourse structure, controlling for other factors such as domain and system knowledge. To do so, we plan to deploy multiple versions of the HR bot. On one end of the discourse structure complexity, we will implement a QA system that takes user input strictly in the format of questions and answers within one-turn. On the other end, we may implement an agent that is able to engage in free-form conversations, including actively initiating and continuing interactions (e.g. asking follow-up questions, changing topics). In the middle of the complexity continuum we may choose systems that handle a limited set of input formats, and does not actively continue conversations. We note that to conduct such studies, it is not necessary to deploy full-fledged working systems. Wizard-of-oz experiments, by having experimenters operating behind the system, would be suitable for the purpose.

Our goal is to empirically identify a set of design priorities for each class of dialogue complexity. For example, previous studies suggested for QA systems to provide suggestion of questions to ask, and to improve perceived quality by providing information source. Meanwhile, we may identify some UX needs specific to complex discourses, such as needs for communication grounding, handling of interleaved dialogues and linguistic phenomena such as anaphora, and engaging in collaborative cognitive processes.

Translating these UX guidelines to system development, it may imply that system modules such as dialogue flow manager and context tracker, conversation patterns such as repair and disambiguation, should be included if the complexity estimator outputs high in the discourse structure dimension. When the output is low (below certain threshold), these modules may not be necessary to balance the development and running costs, but we may include recommended QA design elements such as suggested question lists and information source indicators.

Methodologies for Comparative Studies
We suggest two general directions for conducting comparative studies to derive classified UX guidelines: 1) comparative conversation analysis; and 2) comparative user experiments (A/B testing).

We propose to start with qualitative analyses to identify phenomena and derive patterns of conversational interactions in different complexity classes, and generate hypotheses of UX principles specific to each class. We note that these tasks can be readily supported by methodologies developed in conversation analysis, which deals primarily with identifying regular patterns of conversations and revealing the organized, sometimes tacit reasoning procedures underpinning the behaviors. In fact, comparative approach has long been a pillar of CA. A classic example is the seminal papers on institutional talks [2], which compared conversations in contexts with varying levels of formality, by treating “mundane” conversations in unconstrained everyday life as benchmark against those in institutional contexts, to ceremonial occasions where people speak with pre-arranged contents. Such comparative CA approach was able to reveal what behavioral and sequential features are distinctive about each class of conversations.

Ultimately, our goal is to identify UX principles that are effective for each complexity class. After developing hypotheses of these principles, it is straightforward to test them with experimental studies. For instance, to test the hypothesis that in complex dialogues, it is preferable to include self-repair patterns such as asking clarification questions, we may compare an experimental condition with the agent asking for clarification until reaching confidence, and a control condition without doing it but giving low-confidence output and waiting for user-initiated repairs.

Importantly, the outcome measurements of these experiments need to be given proper considerations. We propose to define a cost-benefit function. The cost may be concerned with development, running costs such as time and space, and practical considerations such as cost of errors. The benefit can be quantified by both objective measurements such as task success and completion time, and subjective measurements such as user satisfaction and trust. While these are both important for conversational UX, they are not necessarily correlated. We also note that, with a large

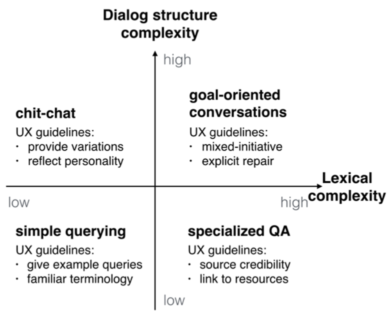

Figure 2: A proposed dialogue complexity scheme and examples of conversational interactions

user base, very fine-grained A\B testing is possible and scalable with automation, opening the possibilities for continuously improved personalization.

A Proposed Dialogue Complexity Scheme

In Figure 2, we propose a complexity scheme based on two dimensions---*lexical complexity* and *discourse complexity*. In linguistics literature, lexical complexity has been measured by lexical richness, variation and sophistication [4], and used to estimate language proficiency and educational development of text authors. We may adopt similar definitions in the expectation to reflect domain specification or user expertise. For example, conversations to solve a medical problem should have higher lexical complexity than day-to-day topics.Discourse complexity, on the other hand, reflects the *richness and pattern complexity of communicative functions between interlocutors*. Although there is a lack of NLP tool to analyze discourse complexity, a potential direction is to develop metrics based on *dialogue acts (DA)*, as DA, by definition, reflects the communicative purpose of conversational input. There are also ongoing efforts in developing DA taggers and computational models [5].

As Figure 2 illustrated, with this simple scheme, we expect to be able to profile conversational interactions in a meaningful way. Next steps would be to conduct comparative studies for these complexity classes, starting from the listed examples as the extreme cases. Desired results from these comparative studies are UX guidelines specific to conversational interactions with each complexity profile, as illustrated in Figure 2. These guidelines can then inform system designs that accommodate individual and contextual differences in conversational interaction styles. Moving forward, other complexity dimensions can be added to enable more fine-grained conversational UX guidelines and adaptive systems.

Conclusion

We propose a novel approach to profile conversational interactions based on dialogue complexity of user input, and identify classified conversational UX guidelines to enable intelligent systems that can adapt to different applications and individual users. We emphasize computation-driven and multi-dimensionality (with correspondence in system modules) in our complexity measures. We propose to conduct comparative conversation analysis and user experiments to derive classified UX guidelines for conversational interactions with different complexity profiles.

References

1. Biber, D. (1992). On the complexity of discourse complexity: A multidimensional analysis. *Discourse Processes*, *15*(2), 133-163.
2. Heritage, J. (2005). Conversation analysis and institutional talk. *Handbook of language and social interaction*, 103-147.
3. Liao, Q. V., Davis, M., Geyer, W., Muller, M., & Shami, N. S. 2016. What Can You Do?: Studying Social-Agent Orientation and Agent Proactive Interactions with an Agent for Employees. In *Proc. DIS '16*, 264-275.
4. Lu, X. (2012). The relationship of lexical richness to the quality of ESL learners' oral narratives. *The Modern Language Journal*, *96*(2), 190-208.
5. Stolcke, A., Ries, K., Coccaro, N., Shriberg, E., Bates, R., Jurafsky, D., ... & Meteer, M. (2000). Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational linguistics*, *26*(3), 339-373.